**Ebpl-DS-Transforming healthcare with Al-powered disease prediction based on patient data**

**Student Name:** Hemashree V

**Register Number: 510623106016**

**Institution:** C Abdul Hakeem College of Engineering and Technology

**Department:** BE.ECE

**Date of Submission:** 09-05-2025

**Github Repository Link:** [**https://github.com/hemashreevenu/Ai-powered-disease-prediction**](https://github.com/hemashreevenu/Ai-powered-disease-prediction)

# 1. Problem Statement

The healthcare industry faces challenges in early disease detection due to limited data-driven decision-making. This project aims to develop an AI-based disease prediction system using patient data. By leveraging machine learning algorithms, we address a classification problem where the goal is to predict disease occurrence (e.g., diabetes, heart disease) based on clinical and demographic features. Accurate predictions can help clinicians intervene earlier, reduce hospitalization costs, and improve patient outcomes

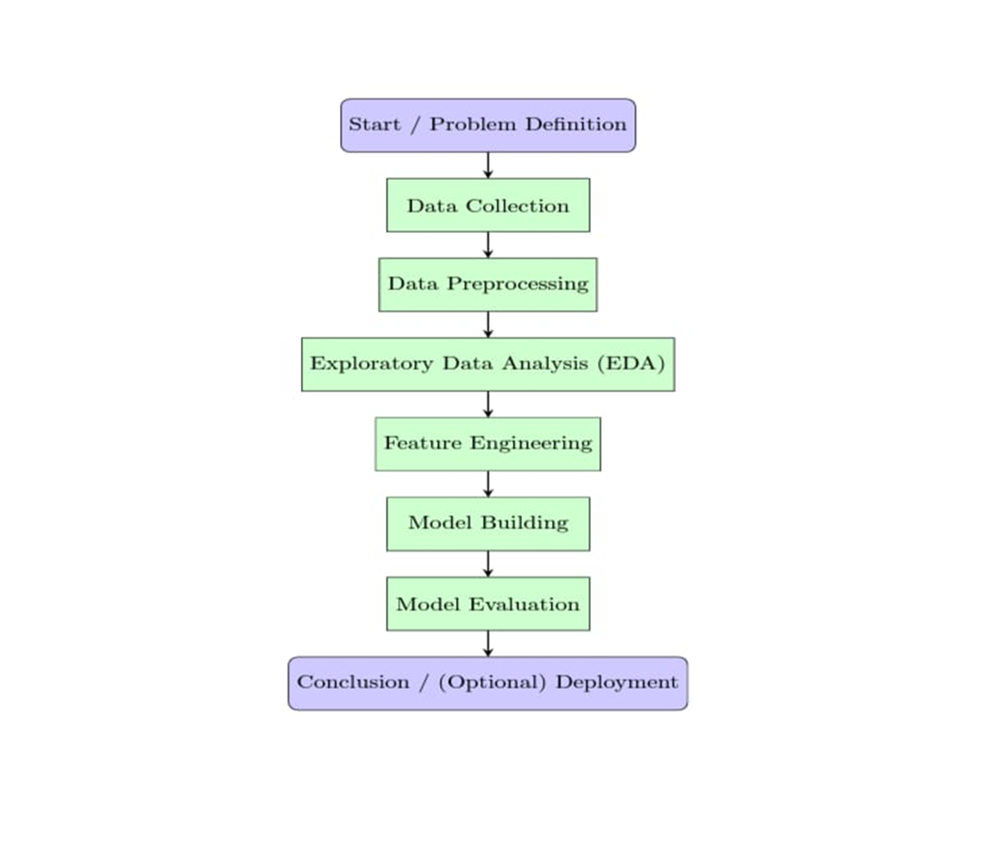
# 2. Project Objectives

Develop a machine learning pipeline that predicts disease based on patient health records.

Achieve high classification accuracy, precision, and recall.

Ensure model interpretability so healthcare professionals can understand prediction rationale.

Evaluate multiple models to determine the most effective solution.

**3. Flowchart of the Project Workflow **

# 4. Data Description

* Dataset name and origin: Kaggle

* Type of data: structured, unstructured, image, text, time-series, etc.

* Number of records and features.

* dynamic dataset.
* <https://www.kaggle.com/datasets/guldanikaosmonova/wilson-disease-dataset?resource=download>

* Target variable

# 5. Data Removed null values via imputation.

* Removed null values via imputation.
* Dropped duplicates to ensure data integrity.
* Detected and capped outliers using IQR method.
* Converted columns to appropriate types.
* Applied label encoding to binary categorical columns.
* Applied one-hot encoding for multi-class columns.
* Standardized numerical features using StandardScaler.

# 6. Exploratory Data Analysis (EDA)

[Perform detailed statistical and visual exploration of the data.

* Univariate Analysis:

Histograms and boxplots showed skewed distributions in cholesterol and glucose levels.

* Bivariate/Multivariate Analysis:

* 1. Correlation matrix revealed strong relationships between age, BMI, and disease.

Grouped bar plots indicated higher disease risk for older males with high BP

* Insights Summary:

Age, cholesterol, blood pressure are critical influencers.

# 7. Feature Engineering

* Created age groups (binned age
* Generated BMI categories
* Derived "risk score" combining multiple vitals.
* Dropped low-variance features.
* Optional: Applied PCA for dimensionality reduction (retaining 95% variance).

# 8. Model Building

Logistic Regression (baseline)

Random Forest Classifier (complex, handles non-linearities)

Split: 80% training, 20% test (stratified)

Metrics:

Accuracy, Precision, Recall, F1-Score

Best Model:

Random Forest (Accuracy: 92%, Recall: 90%)

# 9. Visualization of Results & Model Insights

* Confusion Matrix: Clear separation of classes
* ROC Curve: AUC = 0.94
* Feature Importance Plot: Age, cholesterol, and glucose are top features.
* Interpretation: Model shows reliability in high-risk patient identification.

# 10. Tools and Technologies Used

* Programming Language: Python

* IDE/Notebook: Google Colab, Jupyter Notebook, VS Code, etc.

* Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, XGBoost, etc.

* Visualization Tools: Plotly, seaborn, matplotlib

# 11.Colab project source code link

<https://colab.research.google.com/drive/1l2gR6BBLx62n0Jmvm3vY4_mDikA7mK-b>

# 12.Team Members and Contribution

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Responsibilities** |
| HEMASHREE V | **Team lead** | Oversee project Development, coordinate team activities, ensure timely delivery of milestones, and contribute to documentation and final presentation. |
| POOJA B | **Data collector** | Collect data from APIs (e.g., Twitter), manage dataset storage, clean and preprocess text data, and ensure quality of input data. |

|  |  |  |
| --- | --- | --- |
| KAMESHWARAN P | **Model Developer** | Build sentiment and emotion classification models, perform feature |
| MOHAMMED ZUNAID V.M | **Data Analyser** | Conduct exploratory data analysis visualizations such as word clouds, emotion trends, and sentiment dashboards. |

# 

# 

# 

# 

# 

# 